8

UIUCDCS-R-76-836

A VIDEOTAPE ANALYSIS OF STUDENT PERFORMANCE ON AN INTERACTIVE EXAMINATION

by

Wilfred J. Hansen
Richard Doring
Lawrence R. Whitlock

October 1976

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN · URBANA, ILLINOIS

UIUCDCS-R-76-836

A VIDEOTAPE ANALYSIS OF STUDENT PERFORMANCE ON AN INTERACTIVE EXAMINATION

by

Wilfred J. Hansen
Richard Doring
Lawrence R. Whitlock

October 1976

Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801

This work supported in part by the National Science Foundation under grant EC41511.

TABLE OF CONTENTS

LIST OF FIGURES

1. Introduction

Enlightened design of interactive systems demands insight into the behavior of system users. Such insight is not crucial for users who can be extensively trained, but not all users can receive special training. This paper reports a preliminary study of one variety of untrained users: students taking an examination interactively. In undertaking this study, we had several goals:

i) to explore videotaping as a means to measure and analyze the behavior of terminal users;

ii) to determine why students seemed to be taking as much as twice as long on examinations administered interactively;

iii) to use the results of (ii) to improve our system, and

iv) to stimulate the formation of hypotheses about users and usage of interactive systems.

Among the techniques employed to study terminal users, one obvious approach is to record all terminal input-output activity. This can be accomplished even without system modification by introducing a small processor between computer and terminal and manually examining the results [Treu, 1972]. Alternate approaches to recording the reason for a user's actions are to display a menu of possible reasons [Kulsrud, 1974] and to design the system so various benefits encourage the user to state the nature of the work [Hansen, 1973]. None of these techniques allow detailed examination of the non-terminal activities of the user; for

example, the researcher cannot demarcate reaction to an annoyance and subsequent resumption of productive work. To attempt to solve such problems, we decided to videotape users.

Preliminary results showed that the subjects, on average, did indeed spend longer on interactive versions of examinations (as administered by an early version of the exam system). Using the paper exam time as a base, the additional times included 16% for display generation, 45% for avoidable attributes of the system as it then stood; and, curiously, 37% as a result of longer think times on each problem. However, a "representative times" analysis shows that most subjects spent the same amount of think time on both PLATO and paper.

Section 2 outlines those aspects of the system relevant to this paper. Sections 3 and 4 present the experimental procedure and basic results. Section 5 describes changes made in the exam system which reduce the "PLATO penalty" from 98% to perhaps 20%. Section 6 discusses some observations and hypotheses suggested by our work, especially context-switch time and a theory of nuisance, annoyance, and agitation. The exams, questionnaires, and detailed results of the experiment are presented in a companion paper [Doring, 1976].

Acting on an idea of the first author, the second author conducted the experiment with the aid of the third author. The latter implemented the exam monitor and several problem generators, including two of the four used in the experiment. After the second and third authors coded the tapes and summarized the data, the first author conducted the analyses presented below.

2. Environment

The examination system in question - the Generative Exam System [Whitlock, 1976] - offers many advantages. A wide variety of question schemes are available. Geographic and temporal scheduling problems can be simplified by the network nature of the host system. The system can offer non-passive problems wherein subsequent subproblems reflect prior performance. Each student receives a slightly different form of each question (to encourage honesty). Finally, scores and correct answers are instantly ready for student review, and scores and statistics are instantly ready for the instructor.

The exam system, in turn, is but one component of the Automated Computer Science Education System (ACSES) under development by the Computer Science Department at the University of Illinois [Nievergelt, 1976]. ACSES includes over 100 instructional lessons, some twenty of which are in regular use for over 3000 introductory level students each year. These students constitute the population for which the exam system is designed. They are non-major underclassmen from a wide variety of fields. They have undeveloped typing skills, minimal exposure to interactive systems, and little motivation to learn computer programming.

ACSES is implemented on the PLATO system for computer aided instruction developed by the Computer-based Education Research Laboratory, also at the University of Illinois [Bitzer, 1973]. The University's PLATO system is currently the most fully used; it has 1000 terminals connected and about 400 are usually in operation. From a terminal a student can have access to several thousand hours of instruction, several hundred of which have been polished for - and used in - regular courses.

Though the PLATO system has many advantages, it imposes serious constraints on processing power, memory size, disk accesses, and display speed. Each user is limited to a maximum of ten "TIPS" (thousands of instructions per seconds) and good response cannot be expected if a user requires more than three TIPS. This is adequate for simple question-and-answer interaction, but it cannot support sophisticated versions of data base search, program text analysis, or exam grading. Program and data memories are limited, so large programs and data bases are impractical. Core limitations could be avoided by retrieval from disk, but each user is restricted to an average of one disk access per minute. In the Generative Exam System, this resource is consumed in reading and updating student records for each switch from one problem to another.

PLATO employs a plasma panel terminal. Its 512x512 dot display is high precision but, like a storage tube, cannot support rapid animation. Because it is driven by 1200 baud lines with a maximum of 180 characters per second, the terminal takes ten seconds to display a full page of text. Text may be written in a very rich character set because the terminal provides 128 built-in characters and memory space for another 128 to be defined by the user program. Unfortunately, it can take over ten seconds to transmit the codes to define any substantial number of extra characters.

The keyboard is an augmented typewriter design with a number of "function" keys, including NEXT, BACK, DATA, and HELP. Though a lesson designer can specify any response to these keys, they have certain conventional uses. HELP usually causes display of some additional explanation. BACK moves the lesson back to the last section of material covered. Shifted-DATA often returns the display to the index page for the lesson.

NEXT has two possible meanings: sometimes it terminates an answer and sometimes it simply signals that the student is ready to go on. Usually when the student answers a question and presses NEXT the system responds with "no" or "ok" following the response. This feedback is an important reinforcement during instruction, but must be switched off for exams.

The Generative Exam System is structured as a central monitor and a collection of "problem generator/graders" (PG/G's). The central monitor displays the exam cover page, transfers to individual PG/G's, stores student answers and scores, and generates statistics. Each PG/G displays one or more problems and interacts with the student to accept the answers. The generality of this structure supports an unlimited variety of question types and within each PG/G random generation techniques provide an infinite number of questions.

Each subject took two exams, one form A and one form B. Each form used the same four PG/G's:

1) Arithmetic expressions - 12 variables and their values were displayed and the student was asked to evaluate five expressions involving those variables.

2) FORTRAN syntax - This PG/G included a page of instructions, a cover page, and three problem pages. Each problem page displayed a FORTRAN statement with possibly a syntactic error created with an Extra, Missing, or Replaced basic symbol. For the experiment, form A had assignment, ASSIGN, and PRINT statements, while form B had assignment, GOTO, and DO. In each case the answer was to be given by specifying a type of error (E, M, R, or None) and the associated basic symbols.

3) Print with FORMAT - A print statement and a FORMAT were displayed along with a grid for the answers. The student entered a line of output and then specified where in the grid it should go. Form A had three F format items; form B had three I format items.

4) DO loop - A DO-loop with a PRINT statement was presented and the student had to specify the values that would be printed. This PG/G

graded interactively; each answer was checked when it was entered. Points were immediately deducted for an incorrect answer and a second chance was given. This scheme tried to avoid the problem of propagation of errors.

3. Experiment Design

A two by two design was chosen with subjects in each cell taking two forms of an exam - one on paper and one on PLATO. Both forms were generated by the system and copied on a screen copier to generate the paper version. The design assigned exam forms and methods as shown in figure 3.1. Because the two forms were considered essentially equivalent (and turned out to be so), the subjects were split between PLATO-first and paper-first so neither treatment was favored by the subjects' self-expressed typing ability. Having only one set of videotape equipment and one set of proctors the experiment was run in separate sessions of roughly one hour each.

Subjects were recruited in an elementary computer science course similar to those for which the exam system will be used (Computer Science 101, an introduction to FORTRAN for engineering students, Fall, 1975). They were offered the inducement of practice for the hour exam they would be taking a week later. For fairness, however, the experiment exams were made available to all students after completion of the experiment. (More recently, the exam system has become a popular way to study for exams.) We decided to tape only four subjects for several reasons: the major anticipated effect was large (i.e., slowness on PLATO), the equipment was expensive, and we needed experience. We had enough volunteers to schedule an additional four subjects who were not videotaped; this proved fortunate, because one of the eight subjects

failed to appear at the appointed hour. When students volunteered, they were asked to state their typing ability.

		Method and Form	
Taped Subject	Untaped Subject	First	Second
SA		PL A	pa B
SB	SE	pa A	PL B
SC	SF	PL B	pa A
SD	SG	pa B	PL A

Figure 3.1. Experiment Design. For this paper, the subjects have been randomly coded as SA,..., SG. "PL" = PLATO. "pa" = paper.

Each experiment session began with a practice exam on PLATO, followed by the two exams dictated by the design. The post-test questionnaire elicited reactions to the two methods of giving exams.

The video camera was located slightly behind the subject's shoulder so as to be out of sight while still recording the subject's face and hands, and the general appearance of the PLATO screen or paper exam. Due to low resolution, it was not possible to record the details of the work, so an observer was positioned behind the subject to make a manual record including the sequence in which the problems were worked. A clock was positioned beside the terminal so the time was recorded. To avoid time pressure, the clock was turned so the subject could not see its face.

Data analysis began by coding each "activity" observed into one of the categories listed in figure 3.2.

Brief Title	Code	Description
Think	RT	Read and Think
"	CP	Calculate with Paper and Pencil
Answer	EA	Enter Answers
Select	PS	Problem Selection
Generate	PG	Problem Generation (PLATO only)
Load	LC	Load Character Set (PLATO only)
Display	PP	Problem Presentation (PLATO only)
Trouble	WN	What Next (subject confused)

Figure 3.2. Activity Categories. For most of this paper, CP--calculate with Paper and Pencil--has been lumped together with "Think."

4. Results

Ideally, we would show that our seven subjects mirror the population and that the main result - increased time on PLATO - appeared equally in both taped and untaped subjects. Then it would be reasonable to claim that detailed behavioral analysis of the taped subjects represents the behavior of the entire population. Unfortunately, the subjects do not represent the population and taping did influence behavior. Our task then is to examine closely the ways in which these differences affect the possible conclusions. Fortunately - as we will discuss in section 5 - many effects were obvious from unaided observation, so extensive statistical analysis is not essential.

4.1 Bias Analysis

The essential data for comparing the user population with our seven subjects is presented in figures 4.1 and 4.2. Figure 4.1 shows the backgrounds of the subjects and their times and scores for each experimental treatment. Figure 4.2 shows their responses to the post-test questionnaire.

4.1.1 Do the subjects represent the population?

As far as field of study, typing skill, and approximate hours on PLATO, the subjects are reasonably representative of the population, though perhaps with slightly more years of school. In all cases their ages were consonant with their class standing. Since there were no freshmen, all subjects had prior experience with exams on paper. The four grade dependent variables are all above average for the expected population. This

(a) Background Data

Sub-ject	Year	Field	CS 101 Self-Evaluation	CS 101 Expected Grade	Typing Skill	Approx. Hrs. On PLATO	GPA	First Hour Exam
SA	SO	Cer.Eng.	Good	B	Medium	20	B+	A
SB	JR	Elec.Eng.	Good	B	Slow/sure	3	A-	B
SC	SR	Chem.	Good	B	$\geq$ Good	3	B-	B-
SD	SR	Met.Eng.	Fair-	B	Slow/sure	10	B	B
SE	SO	(LAS)	Fair	C	Slow/sure	10	B	C
SF	JR	Math.	Good	B	Slow/sure	5	B+	B
SG	JR	Physics	Very Good	A	Slow/sure	>>20	A-	A

(b) Results

Sub-ject	First Exam	Total Time (Min) Prac.	Total Time (Min) PLATO	Total Time (Min) Paper	Total Score PLATO	Total Score Paper	Pts/Min PLATO	Pts/Min Paper	Satisfaction
SA	PL A	8	23.2	7.3	25	30	1.1	4.3	1.7
SB	pa A	13	19.8	14.8	24	35	1.2	2.3	3.2
SC	PL B	8	26.7	10.8	27	25	1.0	2.3	3.6
SD	pa B	5	17.9	11.4	38	39	2.1	3.5	3.9
SE	pa A	9	24	14	26	26	1.1	1.9	3.2
SF	PL B	8	16	10	47	36	2.9	3.6	2.4
SG	pa B	11	15	21	49	50	3.3	2.4	3.8
average		8.9	20.4	12.7	33.7	34.4	1.8	2.9	3.1
standard deviation			4.1	4.2	10.1	8.0			.7

Figure 4.1. Summary of all Subjects. For confidentiality, "GPA" and "score on first hour exam" have been converted to letters. Under "first exam", PL = PLATO, pa = paper, A = form, and B = form B. "Satisfaction" is a weighted average of post test questionnaire responses (5 = complete satisfaction with PLATO exam).

		SA	SB	SC	SD	SE	SF	SG
1.	liked PLATO vs. paper (e = much less)	d	c	d	a	c	d	c
2.	experiment bother (c = no bother)	a	c	b	b	c	b	c
3.	PLATO content difficulty (e = too trivial)	c	c	c	c	c	d	d
4.	PLATO instructions difficulty (e = very confusing)	e	c	c	d	c	d	b
5.	concentrate on PLATO (c = not able to)	b	a	a	b	b	a	a
6.	paper content difficulty (e = too trivial)	c	c	c	c	c	d	d
7.	concentrate on paper (c = not able to)	a	a	a	a	a	a	a
8.	PLATO switch frequency (c = once each)	a	b	b	b	c	b	c
9.	PLATO switch difficulty (c = very little)	a	c	c	c	b	b	c
10.	paper switch frequency (c = once each)	b	b	b	c	b	c	c
11.	PLATO keyboard hindrance (e = keyboard much preferred)	b	b	c	e	c	c	a*
12.	PLATO typing impact (d = no influence)	b	c	c	c	c	b	b**
13.	PLATO reading time (e = much less)	b	c	c	b	c	c	c
14.	PLATO vs. paper preference (a = PLATO; b = paper; c = both; d = don't care)	b	d	c	a	a	b	d
15.	did PLATO delay (yes; no)	y	y	n	n	y	y	y
Satisfaction		1.7	3.2	3.6	3.9	3.2	2.4	3.8

Figure 4.2. Summary of Post Test Questionnaire Responses. The taglines in parentheses indicate the permitted range of response and the meaning of that response. *Response "a" meant "no hindrance". "b" through "e" expressed preference. **The subject wrote that the impact was "increased speed and reduced clerical errors".

is reasonable since they volunteered for the sometimes traumatic experience of taking an examination. Unlike some students, the subjects are at least tolerant of exams. Moreover, they are probably more interested in new experiences, computing, PLATO, and interactive systems. Class standing, grades, and interest all tend to bias the experiment away from finding a difference between PLATO and paper because such students can be expected to do better than average at adapting to the new situation of PLATO and the exam system.

Partial evidence for similarity between our subjects and the user population is that the correlation between their GPA and first hour exam scores is .61, a value similar to that for the population.

4.1.2 Were the subjects fairly distributed between treatment groups?

A careful analysis of the differences and similarities between taped and untaped subjects is essential to understand how conclusions from analysis of the tapes apply to the untaped subjects. This section compares the two groups with respect to ability and adaptation to PLATO. Because the groups are reasonably similar, it is not necessary to rigorously control for ability in further analysis.

The initial assignment of subjects to treatment group was based on typing skill (Figure 4.1a) and not score-on-first-hour-exam since that exam did not take place until shortly after the experiment. The first-hour-exam scores are therefore a good check on the assignment of subjects. Examining them we find that among the taped subjects the highest and lowest both started on PLATO and the two in the middle both started on paper. For the untaped subjects, the highest and lowest started on paper and the

middle started on PLATO. In any case, it is difficult to see how this might have influenced the results of the experiment. The similarity of scores between the taped and untaped groups makes comparison reasonable.

A "satisfaction with PLATO" score can be computed by assigning numeric values between one and five to post-test questions 1, 4, 5, 9, and 11-15, with five indicating complete satisfaction. As shown in figure 4.2, these values again suggest that the taped and untaped subjects are sufficiently similar.

Curiously, though without statistical significance, the untaped subjects had a higher average GPA (4.3 vs. 4.1) while the taped subjects had a higher average score on the first hour examination (87 vs. 85). It is possible the taped subjects had more interest in computing relative to other subjects than did the untaped. If so, the assignment of subjects may have introduced a preference for PLATO among the taped subjects, or it may have introduced a propensity for staying on PLATO longer just to "play" with its many bells and whistles. Observation did not seem to bear out either of these possibilities, and they are probably small enough to be overwhelmed by other effects.

The only serious disparity between treatments is that the students who took form A on PLATO averaged nine points higher on the first hour exam than those who took it on paper. As will be discussed below, this disparity does not appear to have had serious consequences.

4.1.3 <u>Did the experimental procedure disturb the students?</u>

There is considerable evidence to show that the experimental situation did influence the behavior of our subjects.

a) Taped and untaped subjects differed markedly in correlations between experiment score and student ability, as measured by grade point average and score on first hour-exam. For the untaped subjects these correlations were nearly 1.0; for the taped subjects they were .14 and -.18. Indeed examination of the scatter plot in figure 4.3 shows a relationship between scores and ability for the untaped, but not for the taped.

b) The time difference between PLATO and paper was much greater for the taped subjects - 10.8 minutes versus 3.3 for the untaped. It seems likely that taped subjects had more trouble due to the pressure they were under.

c) The post-test questionnaire responses showed some negative reaction to the experimental situation. In reply to question two, untaped subjects indicated that the experimental situation was no bother, but more of the taped subjects felt bothered to some extent. Moreover, two taped subjects (and no others) wrote comments on the situation:

"The atmosphere had some affect on concentration."

"Being conscious of being watched and recorded did make me apprehensive....with PLATO <u>you are "broadcasting" your answer to the world</u>."

(underlined in original). The responses to questions four and five are consistent with a negative reaction, but do not themselves indicate one.

Although there are many possible explanations for these results (for example, (a) may have resulted from differing amounts of preparation), the most likely is that subjects were influenced by the experimental procedure. Intimidated by 1) questions on a subject they were just

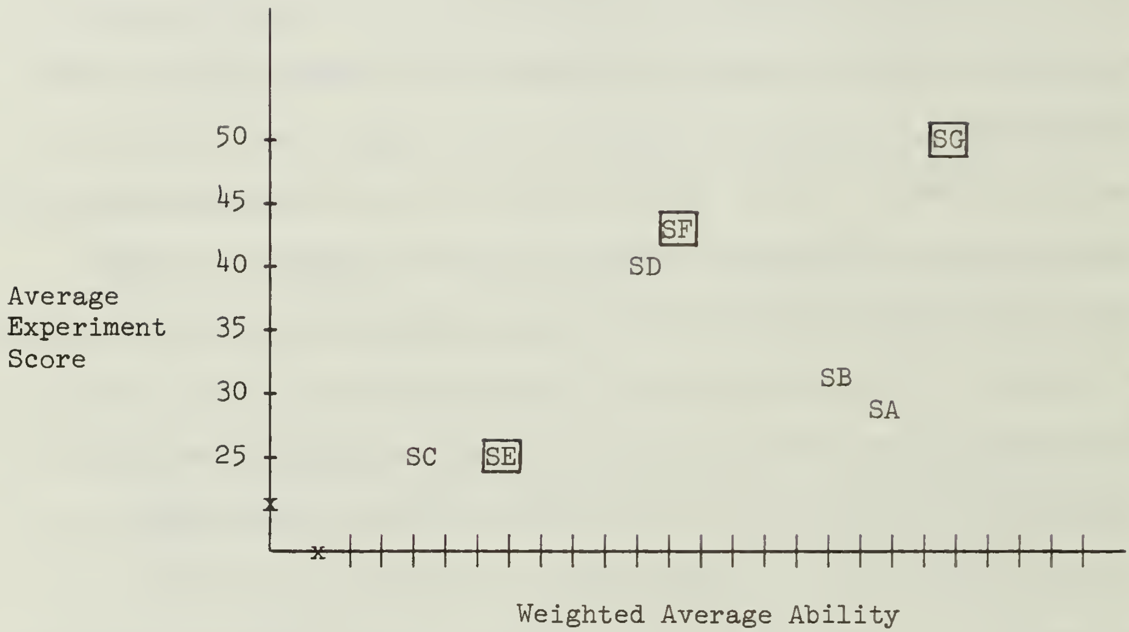

Figure 4.3. Comparison of Ability and Scores. The untaped subject (boxes) showed a relationship (correlation = 1.0-) while the taped subjects did not (correlation = .03). "Weighted Average Ability" is the sum of GPA times 20 and first hour exam score. "Average Experiment Score" is the average of the subject's two scores.

learning, 2) a system they were just learning, and 3) not only close observation, but also preservation for some posterity, the subjects performed erratically. It is our hope that the increased stress did not modify the difficulties but only magnified their visibility.

One technique we should have used more thoroughly to control for apprehension was to observe the untaped subjects more closely. Knowledge that a written record was made of every action would put more pressure on them than our procedure of simply observing which page they turned to.

4.1.4 Were scores and times influenced by exam form or order?

In order to ignore the difference between forms A and B and the difference between first and second treatment, it is necessary to show that these factors did not affect the main result. Non-existence of an effect is impossible to prove, but we have some evidence from Student's t test. Six tests were made with a paired-subjects test for mean different from zero over all seven subjects (df = 6). The following tests all yielded probabilities higher than .4, suggesting they may very well be the result of chance:

the difference in times between form A and form B ($t = -.72$),

the difference in scores between form A and form B ($t = -.52$),

the difference in times between first exam and second exam ($t = .80$),

the difference in scores between paper exam and PLATO exam ($t = .28$).

The difference in times between PLATO and paper was unlikely to have been due to chance ($t = 2.68$, $pr < .05$), as will be discussed in the next section.

The sixth t-test, for the difference in scores depending on order, gave a t value of -1.34 ($pr < .3$), which suggests a closer look. Further examination shows that the group who took PLATO first got about the same score as the group who took PLATO second, while the group who took the paper exam first scored <u>higher</u> than the group who took it second. In fact, all but one subject got a higher score on the first exam and that one was SA who had a very negative reaction to PLATO. These results are counter-intuitive since one would expect subjects to score better on paper and on their second exam. However, detailed examination of the responses to the paper exam do not reveal any systematic explanation. The subjects who took the paper exam second lost most points simply from failure to answer the format question. They also suffered more than the others from confusion over an eight letter identifier generated in an ASSIGN statement (in form A). The failure to improve on the second exam might indicate that there is no learning from an exam, but other research [Kulhavy, 1972] indicates that it simply takes longer for learning from an exam to take effect. More likely is the explanation that subjects became bored or were unable to concentrate on the second exam.

Examination of scores and times by pairs of variables reveals the two interactions shown in figure 4.4a:

1) The group who took form A on PLATO had a higher average score than the other subjects. This result could be expected because this group averaged about ten points higher on the first hour-exam. There were also several respects in which the paper version of form A was probably harder than versions produced interactively. Here, too, the eight letter identifier in the ASSIGN statement was a major factor.

(a) Scores

	PLATO	paper
form A	37.3	30.5
form B	31.0	39.7

	first	second
form A	28.7	37.0
form B	40.8	26.7

(b) Total Time

	first	second
PLATO	22.0	19.2
paper	15.3	9.4

Figure 4.4. Interaction Effects. Note that in each case diagonally opposite values were compiled by the same group of three or four out of the seven subjects.

(a) Average Score. The form-A-PLATO-form-B-paper group was subjects SA, SD, and SG. The form-A-first-form-B-second group was subjects SA, SB, and SE.

(b) Total Time in minutes. The PLATO-first-paper-second group was subjects SA, SC, and SF.

2) The group who started on form A did better than the remaining subjects. Since this result is independent of PLATO vs. paper, student skill, and several other variables, we can propose no reasonable explanation.

A more subtle interaction is revealed by examining total time organized by treatment and order, as shown in figure 4.4b. Those who started on paper took only four minutes longer on PLATO, but the others took much longer on PLATO and much less on paper. One way to express this relation is to say that total time in minutes is

$$9.4 + \{4.4 \text{ if this is the first exam}\} + \{8.2 \text{ if this is a PLATO exam}\} \qquad (4.1.4\text{-}1)$$

The decreased time on the second exam may be from a learning effect or a rush due to loss of interest. The higher scores on the first exam would argue for the second explanation. If this introduces bias, however, it is probably in favor of reducing the time on PLATO and thus is against the finding of the main effect. (The "PLATO exam" effect is partially explained in 4.2.1.)

One clear case of learning effect was the variation in proctoring between early sessions and later. In early sessions we tended to try to have the subjects understand the instructions by reading them just as they would on an examination. When high levels of frustration, as for SA, appeared, we modified our approach to that of quickly helping subjects out of difficulty. In particular, we always showed them how to get out of one of the more difficult spots in the syntax problem. This variation in the experimental procedure should have been avoided by more careful definition of the role of proctors, however, it probably did serve to alleviate some of the pressure of the experimental situation.

4.2 Main Effects

4.2.1 Subjects spent longer taking exams on PLATO

A summary of the data for the four taped subjects is in figure 4.5. Even when combined with the untaped subjects (figure 4.1), the difference in total time (20.4 vs. 12.7) is statistically significant using a paired values t-test for difference not equal to zero ($t = 2.68$, $df = 6$, $pr < .05$). Two different approaches to analysis of the time difference contribute to our understanding of the relative influence of various factors.

The first approach (figure 4.6a) shows that the increased time for taped subjects can be explained as Trouble time and non-productive time:

1) We estimate a "representative PLATO time" by noting that subjects SA, SB, SD, SF, and SG all had times very close to their average of 16.5 minutes after subtracting Trouble time from total time. (We drop the highest and lowest subject. SC spent considerable time in review without changing answers; SE spent 17 minutes on the syntax problem alone, presumably having trouble with the answer scheme.)
2) We calculate a "representative PLATO non-productive time" as the sum of Overhead and Reducible times averaged over the four taped subjects. This value is 4.3 and is close to the value for each subject. (It is reasonable to project this value to untaped subjects because it measures only system actions and its magnitude is not large compared to (1). Incidentally this value explains a large part of the 8.2 minutes PLATO time in equation 4.1.4-1.)
3) The "representative PLATO productive time" is the difference of the values computed in (1) and (2): 12.2 minutes.

times

sub-ject	first	Think		Answer		PLATO Activities			Select		Trouble		Total	
		PL	pa	PL	pa	Generate	Display	Load	PL	pa	PL	pa	PL	pa
SA	PL A	12:07	6:22	1:13	:34	:33	1:17	:20*	2:13	:23	5:28	--	23:11	7:19
SB	pa A	12:27	13:02	1:43	1:29	:31	:43	0*	1:42	:08	2:40	:08	19:46	14:47
SC	PL B	18:43	9:41	:36	:49	:26	1:33	:19*	2:34	:17	2:33	--	26:44	10:47
SD	pa B	10:19	8:55	1:27	1:19	:28	1:34	:18*	2:42	:23	1:04	:44	17:52	11:21
average		13:24	9:30	1:15	1:03	:30	1:17	:19	2:18	:18	2:56	:13	21:53	11:03

activities

SA	PL A	34	26	24	17	3	14	1*	17	9	13	--	106	52
SB	pa A	31	27	22	23	3	8	0*	12	7	8	1	84	58
SC	PL B	39	36	21	17	4	16	1*	20	12	5	--	106	65
SD	pa B	40	22	28	19	4	16	1*	22	7	3	3	114	51

summaries

		pts		Efficiency (pts/prod sec)		Productive (Think+Answer)		PL Overhead (Display+Gen)	Reducible (Load+Select)		#S Activities (Thk,Ans,Sel,Trbl)		# Prob pgs	
		PL	pa	PL	pa	PL	pa	PL	PL	pa	PL	pa	PL	pa
SA	PL A	25	30	.031	.072	13:20	6:56	1:50	2:33	:23	88	52	11	9
SB	pa A	24	35	.028	.040	14:10	14:31	1:14	1:42	:08	73	58	8	6
SC	PL B	27	25	.023	.040	19:19	10:30	1:59	2:53	:17	85	65	13	11
SD	pa B	38	39	.054	.064	11:46	10:14	2:02	3:00	:23	93	51	13	6

Figure 4.5. Summary of Detailed Times for Taped Subjects. Times are in minutes and seconds. *The character set for P2 was preloaded during the practice exam so the times and counts shown are only for review.

(a) "Representative" Times

	PLATO			paper		
	Time	Std. Dev.	Subjects	Time	Std. Dev.	Subjects
1)4) Average of total time less Trouble time	16.5	1.24	(SF,SG,SA, SB,SD)	12.2	1.86	(SE,SF,SB, SC,SD)
2)5) Average non-productive time (Select, Generate, Load, Display)	4.3	.84	(SC,SD,SA, SB)	.3	.12	(SC,SD,SA, SB)
3)6) Representative Productive Time	12.2			11.9		

(b) Distribution of Activities

	average times			
	(1) PLATO	(2) paper	(3) PLATO-paper	(4) (3) as a % of total (2)
Reducible overhead (Load, Select)	2:32	:18	2:14	20%
Inelastic overhead (Generate, Display)	1:46	0	1:46	16%
Productive work (Think, Answer)	14:39	10:33	4:06	37%
Trouble	2:56	:13	2:43	25%
Total	21:53	11:04	10:49	98%

Figure 4.6. Breakdown of the Extra Time Spent on PLATO. For description of the categories see the text. Times are in minutes and seconds. The figures in column (b4) express the excess time on PLATO as a percentage overhead beyond the total time required on paper. (For example, the 4:06 minutes longer productive work on PLATO is 37% of the total 11:04 minutes spent on the paper exam.)

4) The "representative paper time" is total time less Trouble time averaged over subjects SB, SC, SD, SE, and SF, who were all reasonably close to that value: 12.2. (Again we drop the two extreme cases: SA had great Trouble on PLATO but breezed through the paper exam. SG was painstaking enough on paper to achieve a perfect score.)

5) The "representative paper non-productive time" is .3, the average paper problem Selection time for the four taped subjects.

6) The "representative paper productive time" is the difference of the values from (4) and (5): 11.9.

We see that taking account of total time, Trouble time, and non-productive time, the time spent on each exam was about the same.

The second approach provides insight into the relative influence of the factors responsible for the longer times on PLATO. In this approach the videotape data for the four taped subjects are categorized as shown in figure 4.6b. (These times differ from the "representative times" because they include subjects whose times were not representative of all seven subjects.) In the figure, "Productive work" is that time the user spent working on the problem, independent of the method. "Reducible overhead" is time that is easy to avoid by modification of the system. "Inelastic overhead" is inherent in the PLATO system. "Trouble" is that time when the user was having difficulty understanding the requirements. In column (4) the figure expresses PLATO excess time as a percentage of the time the exam took on paper. This basis will permit valid comparison when we discuss reduction of PLATO overhead in section 5.

Several other interesting observations can be made from the data and figure 4.5:

a) All four taped subjects wrote comments about Trouble on the post-test questionnaire. One even remarked - with cautious ambiguity - that, "All of the instructions were not clear."

b) Taped subjects had more "activities" on PLATO than on paper, even when the activities are restricted to those actually performed by the human - Think, Answer, Select, and Trouble. From the data of figure 4.5, we find an average of 65 for PLATO and 57 for paper. Partly this is because subjects reviewed more on PLATO and partly because often when a "Trouble" activity occurred, the subject had to try several times to enter the answer.

c) "Overhead" on the PLATO exam was not negligible; the average time for Display, Generate, and Select per problem page was .36 minutes. Even though comfortingly constant, this time is frustrating to the student because there is little to do but wait. The corresponding value for the paper exam includes only Selection time; an average of .04 minutes. Overhead is especially noticeable on PLATO during review; it averaged over 10% on PLATO and only 3% for those who reviewed on paper. Indeed, for review, Display time is slightly larger than initial presentation because the student's prior work must be displayed; this is offset by the fact that no problem Generation time is needed.

4.2.2 Influences on scores

Factors other than ability influenced scores in this experiment to an unusual extent. Section 4.1.3 has mentioned the impact of the experimental situation; section 4.1.4 discussed the variation due to

exam order. Other factors can be observed in figure 4.7, which presents scores and Think times for each problem.

The figure demonstrates that PLATO itself did not reduce scores. Instead, the major factor in lower PLATO scores is that the interactive grading algorithm in the DO loop problem subtracted too many points for a wrong answer, so subsequent correct answers got no credit. Indeed, we expect that PLATO will actually lead to higher scores because it aids the student in several ways. For example:

SB tried an invalid answer to the syntax problem and was rejected. The retry was correct, so the subject received full credit. (On paper the invalid answer would simply have been marked down.)

SC missed the first line on the DO problem but was given the correct answer and got the rest correct. (Relative grading could have achieved the same result.)

There does not appear to be any inequity in this approach, since all students stand the same chance of being corrected and because invalid answers are likely to be misconceptions that ought to be cleared up on paper by asking the proctor.

Inequity does occur, however, in the level of difficulty of problems. As shown by the average points per minute (last column in 4.7), some problems were fairly easy - problem 2 on PLATO and 4 on paper; while some were too hard - 3 and 4 on PLATO. (Indeed problem 2 on PLATO was so easy that the only points missed were on the ASSIGN statement, a construct not covered in lecture.) Had there been a time constraint, a student who attempted one of the harder problems would be at a disadvantage.

PLATO

	SA (PL A)			SB (pa A)			SC (PL B)			SD (pa B)				
	first try	review	pts	first try	review	pts	first try	review	pts	first try	review	pts	max pts	avg pts min
1. Exprs	2:29	1:03	15	5:06	--	9	4:31	1:27	12	2:25	1:48	9	15	2.4
2. Syntax 1	:22	:07		:13	:06		:43	:39		:16	:22			
" 2	1:30	1:18	10	:48	--		1:15	:23	15	:36	:05	10	15	3.7
" 3	:07	:07		1:35	--		1:33	:29		:15	:47			
3. PRINT	:27	--	0	1:26	:40	0	1:35	1:08	0	2:27	:35	9	10	1.1
4. DO	5:45	:05	0*	4:16	--	0*	5:31	:05	0*	1:59	:11	10	10	.6
total	10:40	2:40	25	13:24	:46	24	15:08	4:11	27	7:58	3:48	38	50	

paper

	SA first try	SA review	SA pts	SB first try	SB review	SB pts	SC first try	SC review	SC pts	SD first try	SD review	SD pts	max pts	avg pts min
1. Exprs	1:57	:16	12	4:01	--	9	3:31	:55	12	3:18	--	6	15	2.8
2. Syntax 1	:20	--		1:28	--		:46	:07		:27	--			
" 2	:19	:14	10	3:08	--	11	1:54	:13	6	:27	--	15	15	2.9
" 3	1:23	:35		:38	--		:56	:09		1:18	--			
3. PRINT	:11	--	0	3:32	--	5	:06	:02	0	1:26	--	10	10	2.8
4. DO	1:41	--	8	1:44	--	10	1:51	--	7	3:18	--	8	10	3.9
total	5:51	1:05	30	14:31	--	35	9:04	1:26	25	10:14	--	39	50	

Figure 4.7. Productive Work Times for each Problem. Times shown are the sum of Think and Answer times in minutes and seconds. Subjects seldom reviewed a question more than once. *Log shows one or more correct answers but grading algorithm subtracted all points.

In addition, there were great disparities in difficulty between different instances of the same problem:

Division problems tend to be harder than other arithmetic expressions, yet a subject might receive 0, 1, or 2 of them.

The syntax generator sometimes inserted no errors in simple statements and sometimes embedded an error deep in a complex statement. It sometimes generated instances of language features not yet studied (as the action of a logical IF).

The format problem required rounding up anywhere from zero to three times.

Sometimes one iteration of the DO loop would modify an array value used by a later iteration. Few students analyzed this correctly.

Careful attention to relative difficulty is essential in interactive exam design. Generators in our improved system are based on a detailed analysis of the factors which contribute to difficulty.

5. System Improvements

As a consequence of this experiement, other observations, and introspection, the exam system has been improved in many ways. This section discusses these improvements roughly in the order a student encounters them. The percentage overheads are taken from figure 4.6; to compare them with improved versions, the denominator is the total time on the paper exam.

5.1 Reducible Overhead

On PLATO, character set Loading and problem Selection time constituted a 20% overhead. Load time has been eliminated by restricting PG/G's to the standard set of 128 characters. It includes almost all characters used in common programming languages, and one or two extra characters can be loaded quickly for special problems (for example, the PL/I logical NOT).

Problem selection time was bloated by a design that required a return to the cover page after every problem. The system has been modified so that the shifted-NEXT key transfers directly to the next problem page and shifted-BACK goes directly to the previous one. These two possibilities account for the vast majority of interproblem transitions. For random access, the shifted-DATA key returns the student to the cover page.

We have not yet found a satisfactory scheme for problems with multiple pages. Certainly shift-NEXT and shift-BACK should move among

the pages of the problem, but should there be a "sub"-cover page for the problem? One pilot PG/G uses a vector of page numbers at the bottom of the page, but this scheme can be a little too confusing. The best interim solution seems to be avoidance of multiple page problems.

5.2 Inelastic Overhead

A 16% overhead resulted from problem Generation and Display, mostly on the latter. Unfortunately, Display time depends on the communication link technology and cannot be reduced by changes to the exam system. Moreover, problem Generation time has increased slightly as more sophisticated approaches have been used. However, several steps have been taken to reorganize Generation and Display so the time required is more palatable and useful:

- Basic problem statements are presented before captions and descriptions of the control keys.
- Problem details are presented as they are generated so the user has something to work on and does not feel the system has halted.
- Detailed instructions that follow standard conventions are accessible via the HELP key, but are not displayed as part of the main display.

Though these techniques create the display efficiently, our observations suggest that users often wait for the complete display before starting to work. Possibly more experience and time pressure will encourage productive use of display construction time.

Other exam systems generate the entire exam when the user first enters the system, or even earlier. We have not chosen this approach because it would be expensive in terms of disk accesses. Two (a read and

a write) would be required for each PG/G during the generation-only phase. It is our hope that generation for each problem can be short enough to keep disruption to a minimum. (Similarly, grading could be deferred until completion of the entire exam, but isn't because of the disk access limitation. Most PG/G's have straightforward, rapid grading algorithms.)

5.3 Productive Time

Although the excess productive time - Thinking and Answering - amounted to 37% of the time to do the exam on paper, there does not seem to be any inherent reason why it should be longer. Indeed, the "representative time" analysis showed similar productive times with both approaches. In addition to the pressures discussed in 4.1.3, longer Think time on PLATO may result from these factors:

(i) Students are used to having an answer judged immediately on PLATO, so they work very hard to be sure they have the correct answer before entering it on the exam.

(ii) They do not know how to change an answer, or perceive it as a difficult process.

(iii) They suspect that a wrong answer will be counted in the grading algorithm. (The suspicion is not unwarranted; we have considered this approach.)

(iv) They fear "exposure" since their answers are more readable by the proctor (and harder to cover up) than they would be on paper.

Improved system design and more practice will help reduce the magnitude of at least the first two of these problems.

The PLATO time to enter an Answer was 19% longer than that on paper, but still amounted to only a second or two more per page. This is

probably not a serious factor. Students did not express difficulty with typing answers and typing ability was not a factor in our results. Many future students will be even more familiar with the keyboard because they will have been using PLATO for other courses.

Nonetheless, the answer entry mechanism is a key element in system design; since no judgement of the answer is expressed, some other action must be taken to show acceptance. In the exam system (even at the time of the experiment) there are two areas for each answer - an entry area and the display area. After entry the answer is moved to the display area, so that even when a new answer is being entered, the old one is still on view in the display area. It is also important that the entry area be adjacent to the display area; this principle is violated by the PRINT FORMAT problem and it is more disconcerting and confusing than the others.

5.4 Trouble Time

Trouble time - a 25% overhead in this experiment - is a very variable quantity and will always exist, even on paper exams; the best remedy is to provide a proctor for every exam. Several of the steps indicated above will also help reduce Trouble time, especially the provision that the shifted control keys for travel between pages must always work. In addition, practice examinations and written pre-test instructions help reduce surprise and confusion.

5. System Improvements (continued)

In view of the above, we can estimate the overhead for doing an exam on PLATO instead of on paper. Reducible time can be eliminated. Inelastic time can be reorganized so the student pays a penalty of perhaps

no more than 5% of the paper exam time. Productive time need not be any longer, especially if students are given practice exams; we hope for a penalty of at most 5-10%. Similarly, Trouble time can be reduced by good design and training, so the penalty should also be no more than 5-10%. We conclude that it is reasonable to expect that the PLATO version of an exam should take no more than 20% longer than the same exam on paper.

An important question is whether students should spend any unnecessary time on exams. The benefits to the instructional staff of reduced preparation and grading time are economically tangible, but does a student receive any benefits for the extra time? Among the benefits are unbiased grading, help with getting the answer into the expected format, immediate knowledge of results, and increased flexibility in temporal or physical scheduling (if the student is willing to risk the absence of a proctor and if the obvious user-identity problem can be solved). We believe these benefits justify the interactive examination approach.

6. Further Analyses

One problem with the videotape technique is the embarrassing richness and variety of data it provides. This section explores a few of the possibilities raised by our data.

6.1 "Context Switch" Time

It is not unlikely that when a subject returns to a familiar page it will take some time to switch mental "context" and recall the material. The experiment offers two types of page return: return to the cover page between problems and return to a problem for review. The "context switch" time on return to the cover page is not directly available in our data because we encoded the entire time from the end of one problem to the start of the next as Select time. However, total problem Select time is known and has four distinct components: Display time for the cover page, "context switch" time, Think time, and the time to press a single key for the next problem. The computation is summarized in figure 6.1. Column (1) is the average Select time for those occasions (including quitting) where selection was via the cover page. Column (2) is the problem Display time computed by dividing the total PLATO Display time by the number of problem pages viewed. (Display time varies slightly with changes in system load.) The next three columns are estimates of the time to choose the next problem and to type the corresponding problem number. Column (3) estimates this as the average time required to select the next syntax problem by

	(1) Select	(2) Display	(3) Shift-NEXT	(4) Answer	(5) Select (paper)	(6) Max of (3, 4)	(7) Think & Switch (1-2-6)
SA	15.8	7.0	1.8	3.0	2.6	3.0	5.8
SB	13.7	5.4	3.0	4.7	1.1	4.7	3.6
SC	14.6	7.2	4.0	1.7	1.4	4.0	3.4
SD	15.3	7.2	4.8	3.1	3.3	4.8	3.3
N	33	45	16	95	35	56	~
Avg.	14.88	6.84	3.56	3.15	2.03	4.00	4.0
SD	7.45	2.38	2.37	<2.30	1.5	2.34	~

Figure 6.1. "Context Switch" Time Calculation. All times are in seconds. "N" is the total number of occurrences for these subjects. The averages and standard deviations are with respect to all occurrences. See text for description of columns. (The value of the 4.0 for the average of column (7) is close to both its horizontal value--horizontally computed from the averages from columns (1), (2), and (6)--or its vertical value--the average of the values in column (7).)

pressing shift-NEXT (the one case of Select time that did not use a cover page). Column (4) is the average time to enter an answer (Answer time divided by number of activities from figure 4.5). For comparison, column (5) gives Select time on paper. Column (6) is just the maximum value from columns (3) and (4). The minimum context switch time plus think time in column (7) is the cover page Select time (1) less the sum of Display time (2) and minimum one-key Answer time (6). The average of 4.0 for context switch plus think time suggests that subjects spent that long just staring at the cover page and recovering from the previous problem. Since the thought involved is negligible, context switch time must be a large fraction of four seconds.

The other approach to context switch time is to consider those occasions when a subject looked at a problem very briefly and went on to another. We can assume they spent that time mostly remembering their work on the problem and that they decided there was little likelihood for improvement. On PLATO, SA had three such occasions, SC had one, and SD had two; for an average time of 6.7 seconds (five were 5 or 7; one was 11). On paper, one subject had four such occasions with an average time of 7.8 seconds (values were 2, 7, 9, and 13).

One implication of "context switch" time is that every time the screen is erased the user pays a time penalty to get reacquainted with the new page. Personal observation suggests that this penalty (though probably less than four seconds) is paid even if the subsequent page has exactly the same format as the former. Thus problem design should emphasize putting a number of similar problems on one display, rather than erasing the screen each time a new question is to be asked.

"Context switch" time suggests that there is some penalty in starting to read any new page and a greater penalty on encountering a new page format. For this reason PG/G's should not use diverse page layouts. In particular, ours all place headings and standard key conventions in the same places.

6.2 PLATO Experience

Intuitively, subjects with more prior exposure to PLATO should do better than others on the automated exam system. However, no hint of this relationship can be derived from our data, whether we consider all seven subjects or only those who were videotaped. For example, we can compare SA and SD, who had a total of thirty hours prior experience, with SC and SB who had a total of six hours. It is true that SA-SD spent less time on the PLATO exam, but they also spent less time on the paper exam; moreover, the ratio of PLATO times is less than the ratio of paper times. I.e., SA-SD were not as much faster as would be predicted by their speed on the paper exam. For scores we can compare the groups on first hour exam score, PLATO score, and paper score. The ratios are 1.07, 1.24, and 1.15, so those with more PLATO experience scored slightly higher on PLATO than would be predicted by other factors alone. However, the difference is not convincing.

It is possible that the learning curve is such that three hours PLATO exposure was enough to learn the features used by the exam system. Another possibility is that prior experience on PLATO was actually detrimental. Prior PLATO experience could not have taught the same set of key conventions as the exam system, especially the return to the cover

page after each problem. Indeed, some PLATO training - for example, waiting for a NO-or-OK judgment on an answer - is antithetical to the behavior of the exam system. Possibly the higher scores achieved by those with more PLATO experience is simply due to more practice with the course material gained by exploiting the available PLATO lessons.

In addition to exposure to PLATO, we can ask whether exposure to typing affects performance at the terminal. One subject who had little PLATO exposure but had had a typing course did reasonably well. Another who claimed to be a better-than-good typist had little Trouble and entered Answers quickly, but had much more Think time on PLATO.

6.3 Review Behavior

A very simple exam system could be implemented if students simply worked exam problems one after the other. No one does, however, so it is important to study the variety and extent of "review behavior" - i.e., of return to problems previously worked on. Such behavior can be a clue to a student's confidence in self and in the answers, so investigation of review behavior can add depth to our knowledge of personality and ability. Two types of review are apparent in our data: "hesitation" - the review of a page prior to going to another, and "rework" - the return to a page.

Our subjects exhibited a variety of review behaviors, as indicated in figures 6.2 and 6.3. For example, SD reworked no problems on paper, but all of them on PLATO. Subject SC - displaying either caution or uncertainty - hesitated on half the problems and reworked almost all. Most subjects hesitated on half the problems, but SC spent considerably more time at it. Our limited data did not reveal any relationships

Subject	Hesitation				Rework						Other Productive work		Scores	
	time		instances		time		instances		change					
	PL	pa	PL	pa	PL	pa	PL	pa	PL	pa	PL	pa	PL	pa
SA (PLA)	:08	:14	1	3	2:40	1:05	5	3	1	2	10:32	5:37	25	30
SB (paA)	:50	:12	3	2	:46	:00	2	0	0	-	12:34	14:19	24	35
SC (PLB)	1:53	:45	3	3	4:11	1:26	6	5	0	1	13:15	8:19	27	25
SD (paB)	1:01	:07	5	1	3:48	:00	6	0	2	-	6:57	10:07	38	39
avg./inst.	:19.3	:08.7			:36.1	:18.9								

Figure 6.2 Work Spent on Review. Times are minutes and seconds spent on Thinking and entering Answers. There were no changes during hesitation because it was defined as that Think time following the last entry of an answer. The last four columns are totals from Figure 4.7.

	Subject	SA	SB	SC	SD	Total
Hesitation	Rework Change	1 0/5 .00	0	4 17/35 .49	2 14/20 .70	7 31/60 .52
	Rework No Change	1 0/10 .00	0	1 12/15 .80	2 19/20 .95	4 31/45 .69
	No Rework	2 13/15 .87	5 29/45 .64	1 7/10 .70	1 10/10 1.00	9 59/80 .73
No Hesitation	Rework Change	2 5/10 50	0	0	0	2 5/10 .50
	Rework No Change	4 37/40 .93	1 0/10 .00	6 16/40 .40	2 5/10 .5	13 58/100 .58
	No Rework	2 0/20 .00	6 30/45 .67	0	5 29/40 .73	13 59/105 .56
	Total	12 55/100 .55	12 59/100 .59	12 52/100 .52	12 77/100 .77	48 243/400 .61

Figure 6.3. Success at Review. A problem is placed in one of these boxes according to the subject's behavior while working it. Thus, the first row contains problems where the subject first hesitated and later changed the answer during a rework. The four figures in each box are the number of problems attempted, the points achieved, the points attempted, and the decimal value of the middle fraction.

between review behavior and any one of score, PLATO satisfaction, or total non-review productive work time.

One good technique for examination taking is to scan the entire exam before starting work. Though none of our subjects adopted this strategy, it is important to note its impracticality on PLATO due to the high overhead of page turning. Despite this, our subjects did rework more problems on PLATO. Such extra rework may reflect uncertainty about the answers, but they were not changed any more than on paper. More likely, the increased rework was to have the assurance that the system really had retained the student's answers.

The data in figure 6.3 illustrate the detail of analysis that can be accomplished with the videotape approach. In this table each problem for each subject is assigned to a cell according to whether the subject hesitated on or reworked the problem and whether the answer was changed during rework. The data suggest the following hypotheses:

a) If a student hesitates and later reworks a problem it is as likely to be changed as not.

b) If there is hesitation, but no rework, the answer is likely to be correct.

c) If there was no hesitation, there is unlikely to be any change during a rework.

Hypothesis (b) might correspond to the subject gaining confidence during the hesitation. Similarly, hypothesis (c) might correspond to a complete confidence in the solution. Such cues, if valid, would provide a basis for evaluation of self-confidence. Subsequent comparison with the final score would provide a measure how realistic the student is. Knowledge

of this factor would be invaluable, for example, in consultation with the student concerning study habits.

6.4 Nuisances, Annoyance, and Agitation

Many of our other observations contribute to the theory of "nuisances, annoyance, and agitation":

A nuisance is a system action or requirement that bothers a user.

Annoyance is the user's immediate response to a nuisance.

Agitation is the cumulative effect of multiple annoyances.

In this theory, annoyance and agitation decay exponentially, but agitation decays more slowly each time one annoyance follows closely on another. Both decay rates depend on the personalities of the individual users. For example, we can hypothesize someone with a "rigid" personality who would adapt well to the constraints of interactive computing, but would have increased agitation when the system fails to be consistent. Even non-rigid personalities may suffer on exams where agitation is compounded with tension and ego-involvement.

Quantification of agitation may have a significant impact on system design. If an economical measure can be found the system could detect problems and modify itself or suggest alternative approaches to the user. If measurement requires extensive subsidiary equipment, it can still be a valuable tool in testing system designs and suggesting training regimens. Ultimately, understanding of the causes of agitation can lead to better methodologies for initial system design.

At present, agitation itself can best be measured indirectly as by the "satisfaction" scale in figure 4.2 or by inference from behavior.

However, direct observation can reveal nuisances such as the three classes we found in our experiment: work habit change, interaction uncertainty, and surprise.

A work habit - for example underlining key words or lifting the page corner while finishing an answer - may help a worker concentrate and reduce anxiety. When a tool change forces a change in habit, the worker will suffer momentary distraction each time a habitual action is thwarted. Unfortunately a switch from paper to an interactive system - no matter how valuable otherwise - requires significant habit changes. The biggest changes are the switch from pen or pencil to keyboard, the increased formality of page turning, and the severe constraints on modification of the visible image. Marginal notes are no longer possible; for even a simple manual calculation the entire torso must turn.

As evidence of changes in work habits we present in figure 6.4 the times subjects spent calculating with paper and pencil. The first two subjects performed similarly on both media, but the other two did much less hand calculation on PLATO. Curiously, the latter two also had higher "satisfaction" scores. In all cases, the total times and scores were similar for both PLATO and paper. Another variety of change was exhibited by SB who pointed at the variables and their values on paper, but not on the PLATO display.

Interaction uncertainty refers to a little noticed disadvantage of the very flexibility which is the greatest advantage of interactive systems. Because a system can exhibit so many behaviors, it far less predictable than a piece of paper; a new user must always wonder, "What

		manual calculations				think and answer				
		PLATO		paper		PLATO		paper		
subject	first	time	#	time	#	time	score	time	score	satisfaction
SA	PLA	1:49	5	1:50	5	1:43	15	:23	12	1.7
SB	paA	:10	1	:23	1	4:56	9	3:38	9	3.2
SC	PLB	:51	3	2:12	6	5:07	12	2:14	12	3.6
SD	paB	:55	2	2:46	1	3:18	9	:32	6	3.9

Figure 6.4. Comparison of Manual Calculation Behavior for the Arithmetic Expressions Problem. Times are in minutes and seconds. "#" is the number of instances of manual calculation. The "productive time" in Figure 4.7 is the sum of the two times shown here. The "satisfaction" score is copied from Figure 4.2.

will it do now?" In an instructional lesson, imaginative variety can help maintain student interest, but on an exam it only heightens anxiety. Uncertainty arises from the variability of the time to create a screen image and from the fact that images may be constructed in random order. The student must integrate the image as it forms (and hope to find and read each piece before the next appears), or wait for the full display. Our system may further increase insecurity by moving each answer after it is entered, and by occasionally rejecting an answer, features which have been cited as advantages in other sections of this paper.

The evidence for interaction uncertainty is sketchy. The videotapes seem to show subjects waiting patiently for the full display and then moving to an attitude of attention when the display is complete. In addition, subjects reviewed more on PLATO, both hesitating and reworking longer and more often (see section 6.3), possibly to check that the system had actually retained their answers.

In section 5 we advocated nonlinear image generation to emphasize key points and utilize generation time. If interaction uncertainty is a severe problem, it may be necessary to reconsider this choice. In any case, students must receive considerable exposure to PLATO and the exam system prior to the exam in order to gain confidence in the predictability of the system.

Surprise nuisances are unusual actions which the student will encounter only rarely, whether because they are bugs or are inconsistent with other system behavior, or only occur when the student performs an uncommon sequence of actions.

Several aspects of the exam system surprised our subjects. The return of the arrow to the top of the arithmetic expression page prompted one subject to turn and ask what had happened. Another subject thoroughly reconsidered a syntax problem after answering correctly, possibly because the advance of the arrow suggested there was another syntax error. The lack of immediate judgment of answers was at variance with all prior PLATO experience. Answer rejection caused many problems since it is completely unlike the behavior of the paper exam. Students had to pause to replace an arithmetic expression answer with an "e" (and "E" wouldn't work either). The DO loop problem not only rejected answers, but only gave one second chance. Worst of all, the syntax problem would reject an answer and then reject all control keys until an acceptable answer was entered. The result of all these answer entry surprises may be a pause by the student after entering each answer. If so, this would be a possible explanation of the hesitation behavior described in section 6.3.

The nuisances of work habit change, interaction uncertainty, and surprise can combine to provoke very negative reactions to a system, as happened to subject SA. Considerable research is needed to learn how to detect and eliminate such problems, preferably before the system is implemented.

7. Conclusions

In videotaping as few as four subjects, it was not our intention to produce statistically defensible results, but rather to discover trends, to help form hypotheses, and to study videotaping as a tool. As a dividend, we were able to find numerous exam system improvements which - while only suggested by the data - were seen to be important by introspection and observation of larger groups of students.

7.1 Results and Action

The primary results in section 4 show that the experiment was not unduly biased by choice of subject or their assignment to groups. Two analyses of the time taken on PLATO versus that on paper were made:

a) "Representative" times constructed by subtracting Trouble time and system overhead from the total time for a representative group showed that the total time for "normal" exams on PLATO was very close to that for paper (12.2 minutes versus 11.9 minutes; the PLATO group had about three minutes more Trouble time and about four minutes more system overhead).

b) Micro-analysis of the eight categories of activities by the videotaped subjects showed that the PLATO group spent more time in each of four groups of activities - Productive work, Easily reducible, Inelastic, and Trouble.

The first of these analyses demonstrated that, under good conditions, PLATO time can be similar to paper time. The second indicated the benefits

to be gained by efforts to reduce times for each group of activities. Among the specific steps we have taken to reduce PLATO time are these, as discussed in section 5.

- Selection time has been eliminated by defining two control keys to move between problem pages (rather than require a trip to the index for every Selection).
- Load time for character sets has been eliminated by prohibiting special character sets within Problem-Generator/Graders.
- Display and Generate time has been masked by reordering displays to present crucial information first and to display as much as possible during each step of generation.
- Trouble and Think time may be reduced by a number of other steps taken to simplify and standardize the various PG/G's.

It is our expectation that as a result of these steps a PLATO exam will usually take no more than 20% longer than a similar paper exam.

7.2 General Model

Through this work we have observed a number of variables operating to influence a subject's performance. One way to organize these diverse factors is sketched in figure 7.1. In this diagram, a student's general personality factors are in the box at the top. The three boxes on the left represent groups of parameters related to specific knowledge of the course material; the three at the right contain the groups of variables relating solely to usage of the terminal and PLATO system. On both the left and right we find some factors inherent in the student, some external, and some derived from interaction of the other two. Finally, at the bottom

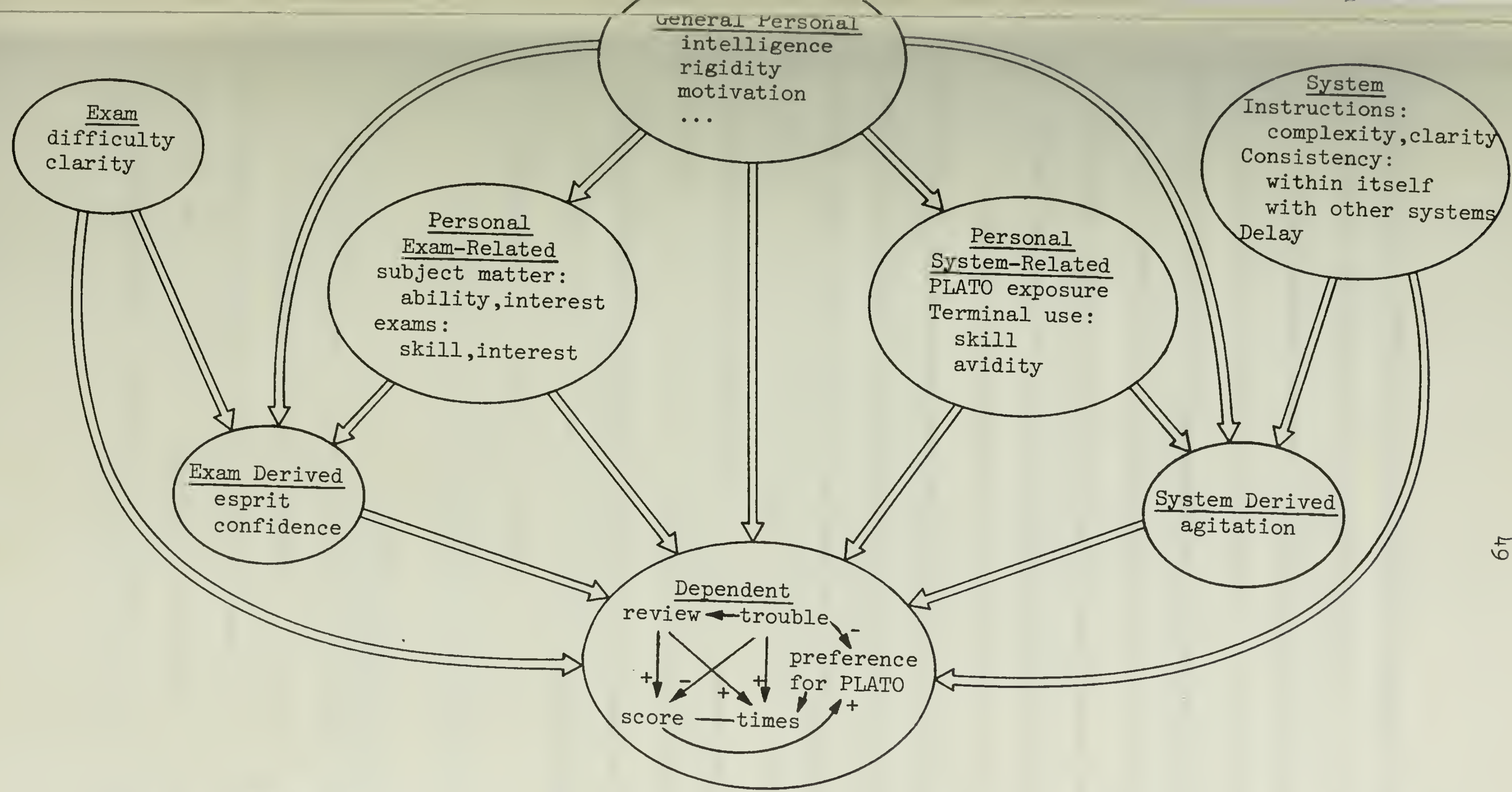

Figure 7.1. Relations Among Groups of Variables Affecting Exam Performance. A double line between two groups indicates that any variable in the first group may affect any in the second. Variables within groups may also interact, as suggested by the inner diagram for the dependent variables. A sign is intended to suggest the probable sign of the correlation.

we find the dependent variables observed in the experiment: score, time, trouble, review behavior, and PLATO preference. The relations between boxes are not simple lines, but sets of lines connecting individual variables, E.G., system-clarity to trouble. Similarly, the variables within each box are not independent; for example probable relations between the dependent variables are shown. The arithmetic signs indicate that trouble may increase time while decreasing score and preference for PLATO. No sign is shown to review because trouble may increase it by increasing uncertainty or decrease it by increasing distaste for the system.

We present the model not as a program for conscientious measurement, but rather as a means of clarifying relationships and suggesting directions for future work. Even partial understanding of a few of the variables can help us

- understand the impact of changes to the exam system and the relative value of different changes.
- understand how knowledge in one area of interactive system design carries over to other areas.
- provide better tools for evaluating students, so as to provide information beyond the mere assignment of a single grade.

7.3 Videotape as a Tool for Experiments

Although videotape has proven its value in numerous fields, it was interesting to see how well it worked for our own experiment. Our expectation of precise measurements was borne out not because we timed the tapes during replay, but because we videotaped a clock. Its face not only showed elapsed time, but also permitted coordination of the tape with the

manual log of the session. Even with the recorded clock as an aid, analysis of three hours of tape took more than 15 hours.

We were pleasantly surprised that audio recording - a trivial by-product of the equipment - was valuable. It enabled us to tell exactly when a key was pressed and to distinguish certain forms of Trouble when the student requested help from the proctor.

A shortcoming we found was that - as we knew in advance - the tapes did not have enough resolution to record the text of the problems and answers. Two possible solutions are to record a close-up of the work with another camera or to record more details manually. More valuable, but more expensive, would be a complete record of terminal input/output, including timings. Such a system might permit computer aided analysis of the terminal session with even more precise timings. In this case videotape analysis would be faster because it would only have to be scanned to observe the user's actions and reactions.

In summary, the videotape technique - when appropriately applied - can be a valuable tool for research into the detailed behavior of interactive computer users.

8. References

Bitzer, Donald L., et al. [1973]
"Computer-Based Science Education," CERL Report X-37, Computer-based Education Research Laboratory, Univ. of Ill, Urbana.

Doring, R., L. R. Whitlock, and W. J. Hansen [1976]
Details of an Experimental Videotape Evaluation of an Interactive Exam System, UIUCDCS-R-76-782, Dept. of Computer Science, Univ. of Ill., Urbana.

Hansen, W. J. [1973]
Design and Evaluation of Systems for Scattered Team Research. Technical Manual TM-5, Dept. of Comp. Sci., Univ. of British Columbia, Vancouver, B.C.

Kulhavy, R. W. and R. C. Anderson [1972]
"The Delay-Retention Effect with Multiple-Choice Tests." J. Ed. Psych V. 63, pp. 505-512.

Kulsrud, H. E. [1974]
"Some Statistics on the Reasons for Compiler Use." Software-Practice and Experience, V. 4, pp. 241-249.

Nievergelt, J., et al. [1976]
ACSES: The Automated Computer Science Education System at the University of Illinois. UIUCDCS-R-76-810, Dept. of Comp. Sci., Univ. of Ill., Urbana.

Treu, S. [1972]
Transparent Stimulation of a Computer User. NBS Report 10 863, National Bureau of Standards, Washington, D.C.

Whitlock, L. R. [1976]
Interactive Test Construction and Administration in the Generative Exam System. UIUCDCS-R-76-821, Dept. of Comp. Sci., Univ. of Ill., Urbana.

IBLIOGRAPHIC DATA HEET	1. Report No. UIUCDCS-R-76-836	2.	3. Recipient's Accession No.

Title and Subtitle

A VIDEOTAPE ANALYSIS OF STUDENT PERFORMANCE ON AN INTERACTIVE EXAMINATION

5. Report Date
October 1976

6.

Author(s)
Wilfred J. Hansen, Richard Doring, Lawrence R. Whitlock

8. Performing Organization Rept. No. UIUCDCS-R-76-836

Performing Organization Name and Address

Department of Computer Science
University of Illinois
Urbana, IL 61801

10. Project/Task/Work Unit No.

11. Contract/Grant No.
NSF EC41511

Sponsoring Organization Name and Address

National Science Foundation
Washington, DC

13. Type of Report & Period Covered
Research

14.

Supplementary Notes

Abstracts

Creful analysis of the behavior of users of interactive systems can yield important isights into the appropriate design of such systems. Because it has not been easy t determine user behavior precisely, we investigated videotaping as a tool. Our aalysis of four students taking examinations both interactively and on paper showed tat they took considerably longer interactively, primarily due to system overhead and touble understanding instructions. The experiment revealed a number of important dsign changes for our system which we expect will reduce the excess time to no more tan 10-20 percent.
Vdeotaping led to analysis of many other variables including context switch time, review behavior and habit changes. These and others observations led us to hypothesize a theory of nuisance, annoyance, and agitation that explains why some students have vry negative reactions to interaction.

7 Key Words and Document Analysis. 17a. Descriptors

Interactive systems
user behavior
videotape
computer aided instruction
nuisance
annoyance
agitation

7b Identifiers/Open-Ended Terms

7c COSATI Field/Group

8. Availability Statement

Unlimited

19. Security Class (This Report) UNCLASSIFIED	21. No. of Pages 57
20. Security Class (This Page UNCLASSIFIED	22. Price